Version dated: February 21, 2017

The effect of non-reversibility on inferring rooted phylogenies

Svetlana Cherlin[1], Tom M. W. Nye[2], Sarah E. Heaps[2], Richard J. Boys[2], Tom A. Williams[3] and T. Martin Embley[4]

[1]*Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, NE1 3BZ, U.K.*

[2]*School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne, NE1 7RU, U.K.*

[3]*School of Earth Sciences, University of Bristol, Bristol, BS8 1RJ, U.K.*

[4]*Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne, NE2 4HH, U.K.*

Abstract.— Most phylogenetic models assume that the evolutionary process is stationary and reversible. As a result, the root of the tree cannot be inferred as part of the analysis because the likelihood of the data does not depend on the position of the root. Yet defining the root of a phylogenetic tree is a key component of phylogenetic inference because it provides a point of reference for polarising ancestor/descendant relationships and therefore interpreting the tree. In this paper we investigate the effect of relaxing the reversibility assumption and allowing the position of the root to be another unknown in the model. We propose two hierarchical models that are centred on a reversible model but perturbed to allow non-reversibility. The models differ in the degree of structure imposed on the perturbations. The analysis is performed in the Bayesian framework using Markov chain Monte Carlo methods. We illustrate the performance of the two non-reversible models in analyses of simulated data sets using two types of topological priors. We then apply the models to a real biological data set, the radiation of polyploid yeasts, for which there is a robust biological opinion about the root position. Finally we apply the models to a second biological data set for which the rooted tree is controversial: the ribosomal tree of life. We compare the two non-reversible models and conclude that both are useful in inferring the position of the root from real biological data sets.

(Keywords: rooting, phylogenetic tree, substitution model)

Introduction

The root of a phylogenetic tree is fundamental to its biological interpretation, providing a critical reference point for polarising ancestor-descendant relationships and for determining the order in which key traits evolved along the tree (Embley and Martin 2006). Despite its importance, most models of sequence evolution are based on homogeneous continuous time Markov processes (CTMPs) that are assumed to be stationary and time-reversible, with the mathematical consequence that the likelihood of a tree does not depend on where it is rooted. Therefore other methods are generally used to root evolutionary trees. The most common approach is to use an outgroup to the clade of interest, or ingroup; the root is then placed on the branch connecting the outgroup to the ingroup (Penny 1976; Huelsenbeck et al. 2002). However, this approach can be problematic if the outgroup is only distantly related to the ingroup because the long branch leading to the outgroup can induce phylogenetic artefacts such as long branch attraction (LBA), potentially interfering with the inference of ingroup relationships and the root position (Felsenstein 1978; Holland et al. 2003; Bergsten 2005). Indeed it has been proposed that the three-domains tree of life, in which Eukaryota represent the sister group to a monophyletic Archaea, could have resulted from LBA (Tourasse and Gouy 1999; Williams et al. 2013). Outgroup rooting is also difficult to apply to the question of rooting the universal tree, for which no obvious outgroup is available. One solution to this problem has been to use pairs of paralogous genes that diverged from each other before the last common ancestor of all cellular life, so that one paralogue can be used to root a tree of the other (Iwabe et al. 1989; Brown and Doolittle 1995; Hashimoto and Hasegawa 1996; Baldauf et al. 1996). However, for any given gene it is difficult to unambiguously establish that duplication took place before the divergence of the domains of life. The number of genes to which this technique can be applied is also limited.

An alternative, but perhaps under-explored, approach to rooting trees is to take a model-based approach, adopting a substitution model in which changing the root position changes the likelihood of the tree. Focusing on homogeneous CTMPs, it is helpful to distinguish between the ideas of *stationarity*, *reversibility* and *homogeneity*. We say that a model is *homogeneous* if it can be characterised by a single instantaneous rate matrix that applies to the whole tree. A homogeneous model is termed *reversible* if the rate matrix can be factorised into a symmetric matrix of exchangeability parameters and a diagonal matrix of stationary probabilities. Similarly we call a rate matrix *reversible* if it permits such a factorisation. Finally a CTMP is *stationary* if the probability of being in each state (e.g. each nucleotide for DNA) does not change over time and the probabilities of transitioning between states over some time interval depend only on the size of that interval and not on its position in time. It follows that all non-stationary models are also non-homogeneous, although the converse need not be true. Models in which one or more of these assumptions is relaxed can give rise to likelihood functions that depend on the position of the root.

For most models that allow root inference, the focus has been relaxing the assumption of homogeneity, typically assigning different reversible rate matrices to different parts of the tree. Generally, these models are non-stationary and allow variation in the

theoretical stationary distribution across the tree. Some also allow variation in the exchangeability parameters (Dutheil and Boussau 2008) although, more commonly, they are fixed over all branches. For example, Yang and Roberts (1995) assigned common exchangeabilities but a different composition vector to each edge of the tree. Heaps et al. (2014) fitted a similar model in a Bayesian framework, but adopted a prior over composition vectors that allowed information to be shared between branches. Whilst biologically persuasive, such non-homogeneous models are, however, highly parameterised and efforts have been made to seek more parsimonious representations. Yang and Roberts (1995) and Foster (2004) both considered models in which composition vectors are applied to groups of edges rather than to a single edge. Blanquart and Lartillot (2006) used a variation of this idea by assuming the compositional shifts occurred according to a Poisson process, independently of speciation events. In the context of nucleotide evolution, Galtier and Gouy (1998) reduced the number of parameters in the model of Yang and Roberts (1995) by using a model parameterised by a single G+C component, rather than three free parameters for the composition vector. But this inevitably came at the cost of a loss of information from the alignment. In a general setting that allowed different reversible or non-reversible rate matrices to be assigned to each edge of the tree, Jayaswal et al. (2011) devised a heuristic to reduce the number of rate matrices using the distances between them as a similarity criteria, and forcing the most similar rate matrices to be identical. However, given the speculative nature of the model search, the algorithm offered no assurance of identifying a global optimum.

In spite of these moves towards parsimony, non-homogeneous models remain substantially more highly parameterised than their homogeneous counterparts. This makes model-fitting computationally challenging, often limiting inference to fixed unrooted trees (e.g. Dutheil and Boussau 2008; Jayaswal et al. 2011) or alignments on a small number of taxa (e.g. Heaps et al. 2014). In this paper we take a Bayesian approach to inference and focus on rooting using a *homogeneous* and stationary, but non-reversible, model that requires only *one* rate matrix. This model has previously been explored by Huelsenbeck et al. (2002), however we build on the work in a number of ways. First, we do not fix the unrooted topology and extend the inferential algorithm to allow inference of rooted trees. This allows us to present a more complete summary of the posterior over root positions and to demonstrate the sensitivity of the analysis to different topological priors. Additionally, whilst Huelsenbeck et al. (2002) only considered small alignments of up to nine taxa, we consider more compelling analyses with data sets of up to 36 taxa. Finally, Huelsenbeck et al. (2002) used a so-called non-informative prior on the rate matrix, with independent uniform distributions for each off-diagonal element. We incorporate prior structure and consider two hierarchical priors that are centred on a standard reversible rate matrix but allow non-reversible perturbations of the individual elements. Our two priors differ in the structure of the perturbation. We test our hierarchical models on simulated data and on a real biological data set for which there is a robust biological opinion about the position of the root. Finally, we apply the models to an open question in biology: the root of the tree of life.

New approaches

Top level model description

We consider a number of aligned homologous sequences and aim to infer the evolutionary relationships among these sequences. These relationships can be described in the form of a bifurcating tree, where each edge represents the period of time over which point mutations accumulate, and each bifurcation represents a speciation event. The nucleotides at each site of a sequence alignment on n taxa can be thought of as independent realisations of a random variable $X = (x_1, ..., x_n)^T$ on a discrete space where $x_i \in \Omega$ and $\Omega = \{\text{A, G, C, T}\}$, for $i = 1, \ldots, n$. The evolutionary process operating along each edge of the tree is described by a homogeneous CTMP, where the future value of a nucleotide at any given site depends on its current value only and does not depend on its past values given this current value, that is

$$\begin{aligned}&\Pr(X(t) = j | X(t_1) = i_1,\ X(t_2) = i_2,\ \ldots,\ X(t_n) = i_n)\\ &= \Pr(X(t) = j | X(t_n) = i_n),\end{aligned}$$

where $t > t_n > t_{n-1} > ... > t_2 > t_1$. The process can therefore be specified by a transition matrix $P(\ell) = \{p_{ij}(\ell)\}$ whose elements $p_{ij}(\ell)$ represent the probabilities of changing from one nucleotide to another over a branch of length ℓ. Equivalently we can represent the process through an instantaneous rate matrix Q, where $P(\ell) = \exp(Q\ell)$. The off-diagonal elements of Q represent an instantaneous rate of change from one nucleotide to another during an infinitesimal period of time. The diagonal elements are specified so that every row sums to zero. If branch lengths need to be expressed in terms of expected number of substitutions per site then the Q matrix has to be rescaled so that $-\sum Q_{ii}\pi_{Q,i} = 1$, where $\boldsymbol{\pi}_Q = (\pi_{Q,A}, \pi_{Q,G}, \pi_{Q,C}, \pi_{Q,T})$ is the theoretical stationary distribution of the process, which can be calculated from Q.

Most phylogenetic models are time-reversible. Reversibility implies that

$$\pi_{Q,i} p_{ij} = \pi_{Q,j} p_{ji}$$

and allows the rate matrix to be represented in the form $Q = S\Pi$, where S is a symmetric matrix containing the exchangeability parameters ρ_{ij}, $i \neq j$, as the off-diagonal elements with $\rho_{ij} = \rho_{ji}$, and $\Pi = \text{diag}(\boldsymbol{\pi}_Q)$ is a diagonal matrix containing the elements of $\boldsymbol{\pi}_Q$. While the reversibility assumption makes statistical models simpler, it has no biological justification, and is applied for computational convenience only. Indeed, there is often evidence of non-reversibility in biological data sets (Squartini and Arndt 2008; Woodhams et al. 2015).

The most common reversible rate matrix, with six exchangeability parameters, is the general time-reversible (GTR) model (Tavaré 1986). The HKY85 model (Hasegawa et al. 1985) is a widely used special case with only two distinct ρ_{ij}, one of which is fixed to

prevent arbitrary rescaling of the Q matrix. The rate matrix Q of this model is then specified by the compositional frequency vector $\boldsymbol{\pi} = (\pi_A, \pi_G, \pi_C, \pi_T)$ and by the transition-transversion rate ratio κ as

$$Q = \begin{pmatrix} \star & \kappa\pi_G & \pi_C & \pi_T \\ \kappa\pi_A & \star & \pi_C & \pi_T \\ \pi_A & \pi_G & \star & \kappa\pi_T \\ \pi_A & \pi_G & \kappa\pi_C & \star \end{pmatrix}.$$

Here the symbol $\star$ is used to indicate that the diagonal elements are specified such that every row sums to zero.

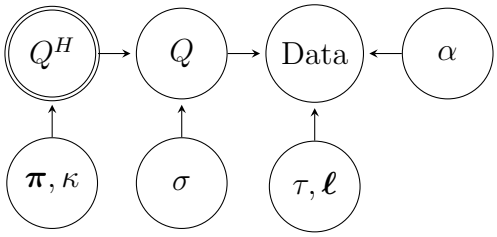
We consider two Bayesian hierarchical models that are both non-reversible and therefore based on an unstructured rate matrix Q. The models differ in the prior they assign to its off-diagonal elements q_{ij}. In each case the prior treats each q_{ij} as a log-normal perturbation of the corresponding element of the unknown rate matrix of a HKY85 model. The first hierarchical model, henceforth called the NR model, utilises one perturbation component, while the more complex model, henceforth called the NR2 model, utilises two perturbation components. The variances of the perturbations are unknown and can provide a measure of the evidence of non-reversibility in the data.

In both models we assume that the variation between the overall rate of substitution events at sites can be modelled by a Gamma distribution with mean equal to 1 (Yang 1993). For computational convenience we approximate the continuous $\mathrm{Ga}(\alpha, \alpha)$ distribution with a discrete $\mathrm{Ga}(\alpha, \alpha)$ distribution with four categories (Yang 1994).

Top level prior distribution

NR model.—

We denote the off-diagonal elements of the rate matrix of the NR model by q_{ij}, and the off-diagonal elements of the rate matrix of the HKY85 model by q^H_{ij}, $i \neq j$, so for instance $q^H_{12} = \kappa\pi_G$. The non-reversibility of the NR model is achieved by a log-normal perturbation of the off-diagonal elements of the rate matrix Q^H using a perturbation component σ as represented in the following directed acyclic graph (DAG):



DAGs are a useful way of representing (especially hierarchical) models graphically. In a DAG, the nodes represent random variables and the directed arrows are used to

indicate the order of conditioning when factorising the joint probability density of all the nodes. A double circle around a node indicates deterministic dependence; in this case Q^H is completely determined once $\boldsymbol{\pi}$ and κ are known. In the DAG above, α is the across-site heterogeneity parameter, τ is the rooted topology and $\boldsymbol{\ell}$ are the branch lengths.

Working element-wise on a log-scale, the off-diagonal elements of the rate matrix of the NR model can be expressed as, for $i \neq j$

$$\log q_{ij} = \log q_{ij}^H + \epsilon_{ij},$$

where the ϵ_{ij} are independent $N(0, \sigma^2)$ quantities. Here the perturbation standard deviation σ represents the extent to which Q departs from a HKY85 structure: the larger its value, the greater the degree of departure. This parameter is treated as an unknown quantity whose value we learn about during the analysis. The unknowns of the hierarchical model therefore comprise: the composition vector $\boldsymbol{\pi}$, the transition-transversion rate ratio κ, the perturbation standard deviation σ, the off-diagonal elements of the rate matrix Q, the shape parameter α, the branch lengths $\boldsymbol{\ell}$ and the rooted topology τ. We express our initial uncertainty about these unknown parameters though a prior distribution that takes the form

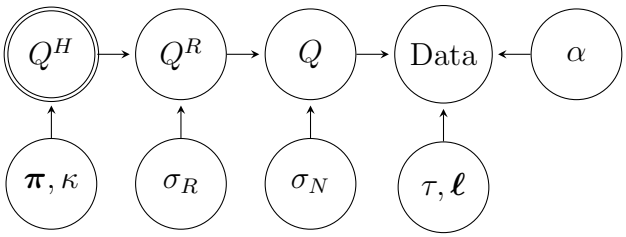
$$\pi(\boldsymbol{\pi}, \kappa, \sigma, Q, \alpha, \boldsymbol{\ell}, \tau) = \pi(Q|\boldsymbol{\pi}, \kappa, \sigma)\pi(\boldsymbol{\pi}, \kappa, \sigma, \alpha, \boldsymbol{\ell}, \tau) \quad (1)$$

in which the top-level prior density $\pi(Q|\boldsymbol{\pi}, \kappa, \sigma)$ has been described above. The bottom level density $\pi(\boldsymbol{\pi}, \kappa, \sigma, \alpha, \boldsymbol{\ell}, \tau)$ will be described in the subsection *Bottom level prior distribution.*

NR2 model.—

Under the NR model, departures from HKY85 structure could lead to a non-reversible model or simply a general time-reversible rate matrix. As such the two types of deviation are confounded and so for any given data set, learning that σ is large does not necessarily provide evidence of non-reversibility. The NR2 model addresses this issue, thereby aiding model interpretation, by using a two-stage process to perturb the underlying HKY85 rate matrix Q^H. The first perturbation is within the space of GTR matrices, perpendicular to the subspace of HKY85 matrices, leading to a reversible rate matrix denoted Q^R. The second perturbation acts on Q^R and is within the space of general rate matrices but perpendicular to the subspace of GTR matrices, leading to a general non-reversible rate matrix denoted Q. These two random perturbations have different variance parameters σ_R^2 and σ_N^2 respectively. Biologically, the variance parameter σ_R^2 represents the extent to which the data contradict the assumption of a common rate of transition and a common rate of transversion. Similarly, the variance parameter σ_N^2 provides a measure of the evidence in the data for the directionality of time.

The general structure of this model can be represented by the following DAG:



The two-stage perturbation procedure is explained further in Appendix 1. The unknown parameters in the NR2 model are therefore: the composition vector $\boldsymbol{\pi}$, the transition-transversion rate ratio κ, the perturbation standard deviation on the reversible plane σ_R, the perturbation standard deviation on the non-reversible plane σ_N, the shape parameter α, the branch lengths $\boldsymbol{\ell}$ and the rooted topology τ. We also have latent variables comprising $\nu_1, \ldots, \nu_5$ for the reversible perturbation, and $\eta_1,\ \eta_2,\ \eta_3$ for the non-reversible perturbation (see Appendix 1). The prior distribution of these unknowns takes the form

$$\begin{aligned}&\pi(\boldsymbol{\pi}, \kappa, \sigma_R, \sigma_N, \boldsymbol{\nu}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)\\ &\quad = \pi(\boldsymbol{\nu}|\sigma_R)\pi(\boldsymbol{\eta}|\sigma_N)\pi(\boldsymbol{\pi}, \kappa, \sigma_R, \sigma_N, \alpha, \boldsymbol{\ell}, \tau).\end{aligned} \tag{2}$$

where the top-level prior distributions with densities $\pi(\boldsymbol{\nu}|\sigma_R)$ and $\pi(\boldsymbol{\eta}|\sigma_N)$ are $\nu_i \sim N(0, \sigma_R^2)$ for $i = 1, \ldots, 5$ independently, and $\eta_i \sim N(0, \sigma_N^2)$ for $i = 1,\ 2,\ 3$ independently (see Appendix 1). The bottom level density $\pi(\boldsymbol{\pi}, \kappa, \sigma_R, \sigma_N, \alpha, \boldsymbol{\ell}, \tau)$ will be described in the following subsection.

Bottom level prior distribution

NR model.— The bottom-level prior density $\pi(\boldsymbol{\pi}, \kappa, \sigma, \alpha, \boldsymbol{\ell}, \tau)$ from (1) takes the form

$$\pi(\boldsymbol{\pi}, \kappa, \sigma, \alpha, \boldsymbol{\ell}, \tau) = \pi(\boldsymbol{\pi})\pi(\kappa)\pi(\sigma)\pi(\alpha)\pi(\boldsymbol{\ell})\pi(\tau)$$

to reflect our initial assessment of independence between these parameter blocks.

The composition vector $\boldsymbol{\pi}$ is defined on the four-dimensional simplex, that is, it has four positive elements, constrained to sum to one. We choose to assign it a Dirichlet prior, $\boldsymbol{\pi} \sim \mathcal{D}(a_\pi \boldsymbol{\pi}_0)$, where $\boldsymbol{\pi}_0 = (0.25, 0.25, 0.25, 0.25)$ is the mean and a_π is a concentration parameter (we take $a_\pi = 4$). This prior is exchangeable with respect to the nucleotide labels. We adopt a log-normal prior for the transition-transversion rate ratio $\kappa \sim LN(\log \kappa_0, \nu^2)$, where $\kappa_0 = 1$ and $\nu = 0.8$. The parameters of the prior for κ represent our belief that the probability of κ exceeding 2 is 0.2, i.e $\Pr(\kappa < 2) = 0.8$. The perturbation parameter σ is assigned an Exponential prior $\sigma \sim \text{Exp}(\gamma)$, where the rate $\gamma = 2.3$ reflects our prior belief that the probability of σ exceeding 1 is 0.1, i.e $\Pr(\sigma < 1) = 0.9$. Together with the rest of our hierarchical specification, this choice induces a prior for the stationary distribution $\boldsymbol{\pi}_Q$ in which little density is assigned to vectors where some characters are heavily favoured over the others.

The branch lengths are assigned independent Exponential priors $\ell_i \sim \text{Exp}(\mu)$, where $i = 1, ..., k$ and k is the number of edges. The rate μ equals 10, so that $\text{E}(\ell_i) = 0.1$. The shape parameter α is assigned a Gamma prior, $\alpha \sim \text{Ga}(10, 10)$, which ensures the expected substitution rate in the $\text{Ga}(\alpha, \alpha)$ model for site-specific substitution rates is modestly concentrated around 1.

We define a *root type* as the number of species on each side of the root. For example, the root type $1 : (n-1)$ represents a root split on a pendant edge, $2 : (n-2)$ represents a root split between two taxa and all others, etc. A uniform prior over rooted topologies assigns a prior probability of more than 0.5 to root splits of the type $1 : (n-1)$, in other words, to roots on pendant edges. We felt that deeper roots are generally more biologically plausible and should be assigned higher prior mass, whilst still retaining a diffuse initial distribution. We therefore chose to assign the rooted topology a prior according to the Yule model of speciation, which assumes that at any given time each of the species is equally likely to undergo a speciation event. This generates a biologically defensible prior in which all root types receive the same prior probability if n is odd, and a near uniform distribution if n is even, but with $n/2 : n/2$ root types receiving half the prior probability of the other root types. The probability of generating a n-species tree T under the Yule distribution is calculated by dividing the number of labelled histories for the tree T by the total number of all possible labelled histories on n species (Steel and McKenzie 2001). This probability depends on the complete rooted topology and therefore has to be re-calculated at every iteration of the Metropolis Hastings algorithm used for inference. To save computational time, we therefore additionally introduce an approximation to the Yule prior, which we term the *structured uniform prior*, that assigns equal prior probability to all root types. In order to sample a rooted topology from this distribution we first sample a root type uniformly. We then sample uniformly from the set of all rooted topologies with that root type. Computationally, this prior is more convenient than the Yule prior because its mass function is independent of the particular unrooted topology and only considers the root split. It also has the advantage of being uniform on root types for all n. Posterior sensitivity to the choice of topological prior will be discussed in the *Analysis of experimental data* subsection.

NR2 model.—

The bottom-level prior density $\pi(\boldsymbol{\pi}, \kappa, \sigma_R, \sigma_N, \alpha, \boldsymbol{\ell}, \tau)$ from (2) takes the form

$$\begin{aligned}&\pi(\boldsymbol{\pi}, \kappa, \sigma_R, \sigma_N, \alpha, \boldsymbol{\ell}, \tau)\\ &\quad = \pi(\boldsymbol{\pi})\pi(\kappa)\pi(\sigma_R)\pi(\sigma_N)\pi(\alpha)\pi(\boldsymbol{\ell})\pi(\tau).\end{aligned}$$

The rate heterogeneity parameter α, branch lengths $\boldsymbol{\ell}$, rooted topology τ and the parameters $\boldsymbol{\pi}$ and κ of the reversible Q^H matrix are assigned the same priors as those used for the NR model. Both perturbation standard deviations are assigned the same prior as their analogue, σ, in the NR model, i.e. $\sigma_R \sim \text{Exp}(2.3)$ and $\sigma_N \sim \text{Exp}(2.3)$.

Results

Taking a Bayesian approach to inference, we fitted the NR and NR2 models to the data sets described in this section using a Markov chain Monte Carlo algorithm. Full details of the inferential procedure are provided in the *Materials and methods* section.

Analysis of simulated data

Our simulations aim to explore two aspects: (i) the effect of different levels of non-reversibility in the data on root inference; (ii) the effect of different topologies and branch lengths on root inference.

Different levels of non-reversibility in the data.—

Here we explore the posterior when the NR and NR2 models are fitted to simulated data that contain different levels of non-reversibility. The tree used to simulate the data is a random 30-taxon tree (generated under the Yule birth process), with the branch lengths simulated from Ga(2,20). The lengths of the branches adjacent to the root are simulated from Ga(1,20) such that the combined length of these two branches corresponds to a Ga(2,20) random variable (Supplementary Fig. 1). In order to simulate the alignments, we first fix the underlying reversible HKY85 rate matrix (Q^H matrix) using the values $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)$ and $\kappa = 2$. We then apply different types of perturbation to the Q^H matrix.

NR model. Five different values of the perturbation standard deviation σ were used to simulate the data: $\sigma = 0$, 0.05, 0.1, 0.2, 0.3. For each value of σ nine different data sets of length 2000 bp were simulated, the first five having different rate matrices (data sets 1 - 5), and the last five having the same rate matrix (data sets 5 - 9). Thus the former five data sets have different stationary distributions $\boldsymbol{\pi}_Q$, while the latter five data sets have the same stationary distribution. This type of alignment simulation allows us to investigate different sources of variability in the data. All the alignments were simulated using a gamma shape heterogeneity parameter generated from $\text{Ga}(10, 10)$. Note that the case of $\sigma = 0$ corresponds to the reversible HKY85 model. The other values of σ were chosen so that the prior for the stationary distribution induced by the log-normal perturbation would be in the range of values estimated for real data; as σ increases, significant support is given to highly biased compositions, and for $\sigma > 0.3$ these are biologically unrealistic (Supplementary Fig. 2).

To provide a consistent measure of non-reversibility across both the NR and NR2 models, we consider the value of Huelsenbeck's I statistic ($I = \sum_{ij} |\pi_i q_{ij} - \pi_j q_{ji}|$, Huelsenbeck et al. (2002)). Under a reversible model, $\pi_i q_{ij} = \pi_j q_{ji}$ for all $i \neq j$, and so $I = 0$. However, I is strictly positive for non-reversible models, with larger values indicating a greater degree of non-reversibility. The values of Huelsenbeck's I statistic for the models used to generate the data in these experiments are shown in Table 1.

Table 1: Values of Huelsenbeck's I statistic for the Q matrices used in the simulations with the NR model. By design, there is a strong positive correlation between σ and I.

Data Set	$\sigma = 0$	$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.3$
1	0.0000	0.0398	0.0534	0.1281	0.3191
2	0.0000	0.0287	0.1428	0.2553	0.1483
3	0.0000	0.0407	0.0438	0.1105	0.0991
4	0.0000	0.0500	0.1163	0.1292	0.1383
5-9	0.0000	0.0628	0.0322	0.0827	0.1088

Table 2: Marginal posterior probabilities of the correct root split for the simulations with the NR model and the Yule prior. The posterior means for Huelsenbeck's I statistic are indicated in parentheses. When the correct root split is a modal root split, the corresponding marginal posterior probability appears in bold.

Data Set	$\sigma = 0$	$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.3$
1	0.08 (0.02)	0.09 (0.02)	0.10 (0.07)	**0.40** (0.16)	**0.88** (0.30)
2	0.09 (0.04)	0.12 (0.03)	0.19 (0.13)	0.02 (0.26)	**0.44** (0.19)
3	0.10 (0.03)	0.10 (0.03)	0.07 (0.04)	0.03 (0.11)	0.22 (0.12)
4	0.07 (0.02)	0.07 (0.07)	0.17 (0.12)	**0.23** (0.08)	0.16 (0.15)
5	0.08 (0.04)	0.17 (0.05)	0.04 (0.05)	0.09 (0.08)	**0.32** (0.10)
6	0.08 (0.02)	**0.21** (0.05)	0.06 (0.02)	0.13 (0.08)	**0.50** (0.14)
7	0.09 (0.02)	**0.21** (0.06)	0.08 (0.01)	**0.21** (0.09)	0.10 (0.11)
8	0.08 (0.04)	0.23 (0.07)	0.06 (0.03)	0.16 (0.10)	0.03 (0.08)
9	0.10 (0.03)	0.23 (0.04)	0.11 (0.04)	0.16 (0.08)	0.14 (0.10)

Table 2 summarises the marginal posterior probabilities of the correct root split and the posterior means for Huelsenbeck's I statistic for the data simulated with the NR model and analysed under the Yule prior (the posterior distributions of the root splits are shown in Supplementary Fig. 3). When $\sigma = 0$ the posterior of the root splits is identical to the prior (not shown) because the data contain no information about the root. As σ increases, the root is often inferred better, with $\sigma = 0.3$ demonstrating the best root inference of all analysed values of σ. However, the analyses of nine simulated data sets for each value of σ do not show identical behaviour. There is substantial variability between the data sets, even those simulated with the same rate matrix, and the true root split is not inferred well in all experiments. The true unrooted topology, however, is inferred with posterior probability close to one in all cases (Supplementary Fig. 4). This suggests that in addition to inferring the unrooted topology, we can also use the NR model to extract some information about the root. Moreover, as expected, the greater the degree of non-reversibility, the stronger the signal from the data.

In order to evaluate the sensitivity of the analysis to the topological prior, the same

Table 3: Values of Huelsenbeck's I statistic for the Q matrices used in the simulations with the NR2 model. By design, there is a strong positive correlation between σ_N and I.

Data Set	$\sigma_N = 0$	$\sigma_N = 0.1$	$\sigma_N = 0.25$	$\sigma_N = 0.5$	$\sigma_N = 1.0$
1	0.0000	0.0550	0.2327	0.3282	1.0416
2	0.0000	0.0366	0.1871	0.4423	0.9019
3	0.0000	0.0737	0.3297	0.4699	0.7494
4	0.0000	0.0538	0.1675	0.3654	0.7282
5-9	0.0000	0.1012	0.3541	0.4402	0.9948

analysis was performed using the structured uniform prior (Supplementary Tab. 1, Supplementary Fig. 5 and 6). This analysis gave very similar results, as we might expect given the similarity between the two priors.

NR2 model. The simulations were performed in a similar manner as for the NR model. Nine alignments were created for each of five values of $\sigma_N = 0$, 0.1, 0.25, 0.5, 1.0. In all the simulations we used the same value for the reversible perturbation, $\sigma_R = 0.1$. Note that the case of $\sigma_N = 0$ corresponds to the GTR model. The values of $\sigma_N = 0.1$, 0.25, 0.5, 1.0 were chosen so that in the prior for the stationary distribution, some nucleotides are not heavily favoured over the others (Supplementary Fig. 7). We note that this type of perturbation allows us to use larger values of σ_N in comparison to the values of σ in the NR model, while still maintaining a realistic stationary distribution. This, in turn, means we can simulate data from models with a greater degree of non-reversibility and, correspondingly, larger values of Huelsenbeck's I statistic. This is illustrated in Table 3 which displays the values of Huelsenbeck's I statistic for the models used to generate the data in these experiments. As for the NR model, for each value of σ_N the first five alignments were simulated from different rate matrices (data sets 1 - 5), while the last five alignments were simulated from the same rate matrix (data sets 5 - 9). All the alignments were simulated using a gamma shape heterogeneity parameter simulated from Ga(10, 10).

Table 4 summarises the marginal posterior probabilities of the correct root split and the posterior means for Huelsenbeck's I statistic for the data simulated with the NR2 model and analysed under the Yule prior (the posterior distributions of the root splits are shown in Supplementary Fig. 8). As with the NR model, when $\sigma_N = 0$ the posterior probability of the root splits is very similar to the prior (not shown). This is because the data contain no information about the root position when simulated under a reversible model. As σ_N increases, the root is inferred better, with $\sigma_N = 1$ demonstrating the best root inference of all the values of σ_N analysed. For the simulations under the NR2 model, the posteriors are more concentrated around the true root position than they had been for the simulations under the NR model. However, comparing the values for Huelsenbeck's I statistic in Tables 1 and 3, this is simply because the data simulated under the NR2 model generally had a higher degree of non-reversibility. Indeed, when fitting the NR model to

Table 4: Marginal posterior probabilities of the correct root split for the simulations with the NR2 model and the Yule prior. The posterior means for Huelsenbeck's I statistic are indicated in parentheses. When the correct root split is a modal root split, the corresponding marginal posterior probability appears in bold.

Data Set	$\sigma_N = 0$	$\sigma_N = 0.1$	$\sigma_N = 0.25$	$\sigma_N = 0.5$	$\sigma_N = 1.0$
1	0.07 (0.24)	0.11 (0.04)	**0.63** (0.23)	**0.92** (0.35)	**0.99** (1.03)
2	0.09 (0.25)	0.08 (0.03)	0.07 (0.16)	**0.87** (0.41)	**0.95** (0.76)
3	0.05 (0.05)	0.13 (0.08)	0.20 (0.36)	0.29 (0.49)	**0.98** (0.69)
4	0.06 (0.07)	0.07 (0.03)	0.09 (0.21)	**0.63** (0.33)	**0.99** (0.77)
5	0.08 (0.02)	0.22 (0.10)	**0.34** (0.34)	**0.91** (0.51)	**1.00** (1.03)
6	0.07 (0.02)	0.13 (0.04)	0.21 (0.37)	**0.92** (0.51)	**0.99** (1.00)
7	0.07 (0.02)	0.03 (0.13)	**0.48** (0.32)	**0.88** (0.46)	**0.95** (0.94)
8	0.08 (0.03)	0.18 (0.08)	**0.36** (0.36)	**0.97** (0.45)	**0.99** (1.02)
9	0.08 (0.02)	0.09 (0.07)	0.23 (0.32)	**0.65** (0.44)	**0.99** (0.98)

the data simulated under the NR2 model, we obtained very similar root inferences to those summarised in Table 4, with strong posterior support for the correct root position for large σ_N.

In terms of inference for the unrooted tree, the true topology had posterior probability close to 1 in all cases (Supplementary Fig. 9). The analysis of the same data sets performed with the structured uniform prior showed similar results (Supplementary Tab. 2, Supplementary Fig. 10 and 11).

Different topologies and branch lengths.—

In a Bayesian analysis, the posterior distribution reflects information from both the prior and the data. When the prior and likelihood are comparably concentrated, but in conflict, the posterior can only represent a middle ground. In phylogenetics, inferences can be highly sensitive to the choice of prior for branch lengths and the topology itself (Yang and Rannala 2005; Alfaro and Holder 2006).

Motivated by the kinds of conflicts that are likely to arise in the analysis of real biological data, we consider the robustness of posterior root inferences to conflicting prior and likelihood information concerning the rooted topology and branch lengths. In our analyses we adopt the commonly used Exp(10) prior for branch lengths and a Yule prior (or the approximating stuctured uniform prior) over rooted topologies. An Exp(10) prior for branch lengths asserts a strong prior belief that edges will be reasonably short. Therefore, given an unrooted topology that contains a long branch, the prior will typically support placement of the root midway along this branch in order to break it up into two shorter ones. The Yule prior for rooted topologies assigns a (near) uniform distribution to all root types. However, there are generally many more trees of unbalanced types, like $1 : n-1$, than there are of more balanced types like $n/2 : n/2$ for n even or $(n-1)/2 : (n+1)/2$ for n odd. It follows that a topology that is more balanced will typically receive more prior mass

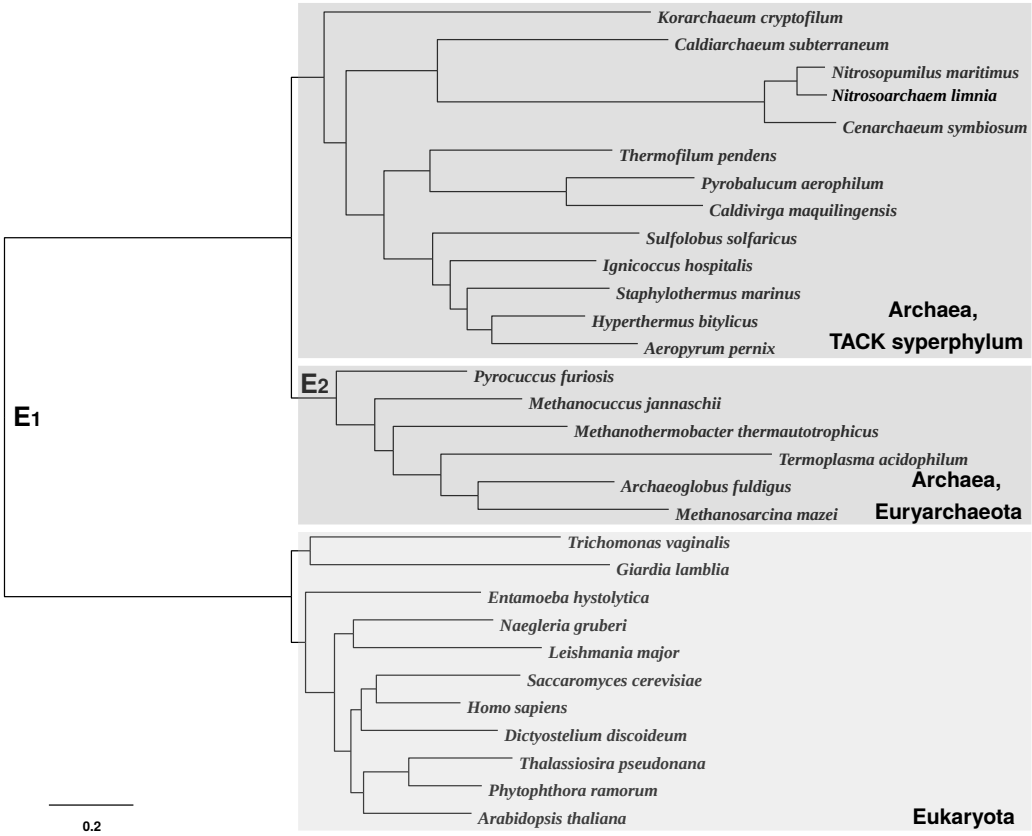


Figure 1: An unrooted 30-taxon tree derived from a recent analysis (Williams et al. 2012) describing the relationships between Archaea and Eukaryota. A root on the branch E_1 corresponds to the three-domains hypothesis (located between monophyletic Archaea and Eukaryota), while a root on the branch E_2 corresponds to the eocyte hypothesis (located within paraphyletic Archaea, separating Euryarchaeota from the clade comprising the TACK superphylum and Eukaryota).

than a topology that is more unbalanced. In the remainder of this subsection we therefore use simulation to examine posterior robustness in cases where prior-likelihood conflict arises due to a data generating tree that is unbalanced or that contains a long branch.

We base our simulations on an unrooted 30-taxon tree derived from a recent analysis (Figure 1) (Williams et al. 2012). This tree describes the relationships between Archaea and Eukaryota. These relationships are still debated, concentrating on two competing hypotheses about the tree of life: (i) the three-domains hypothesis, according to which the root of the tree comprising Archaea and Eukaryota is placed on the branch separating monophyletic Archaea from monophyletic Eukaryota (branch E_1), and (ii) the eocyte hypothesis which places the root within a paraphyletic Archaea (branch E_2). Based on this unrooted tree, we construct six different rooted trees by changing the placement of the root and the length of the branch E_1 according to Table 5.

Table 5: Six rooted trees for simulating the data. The trees have the unrooted topology of the tree depicted in Figure 1 but differ in the placement of the root and the length of the branch E_1. Note that if a tree is rooted on branch E_i, the root is placed at the middle of E_i.

Tree	Root edge	Length of E_1
1	E_1	1.3
2	E_2	1.3
3	E_1	0.1
4	E_2	0.1
5	E_1	0.3
6	E_2	0.3

Trees 1, 3 and 5 are fairly balanced with root type $11:19$, whilst Trees 2, 4 and 6 are more unbalanced with root type $6:24$. The Yule prior assigns almost 30% more mass to the former rooted topology. In Trees 1 and 2 and, to a lesser extent, Trees 5 and 6, the unrooted topology contains a long internal branch. In Trees 3 and 4 this internal branch is short. Given the unrooted tree depicted in Figure 1, the prior will therefore support placement of the root on branch E_1, particularly if this branch is long.

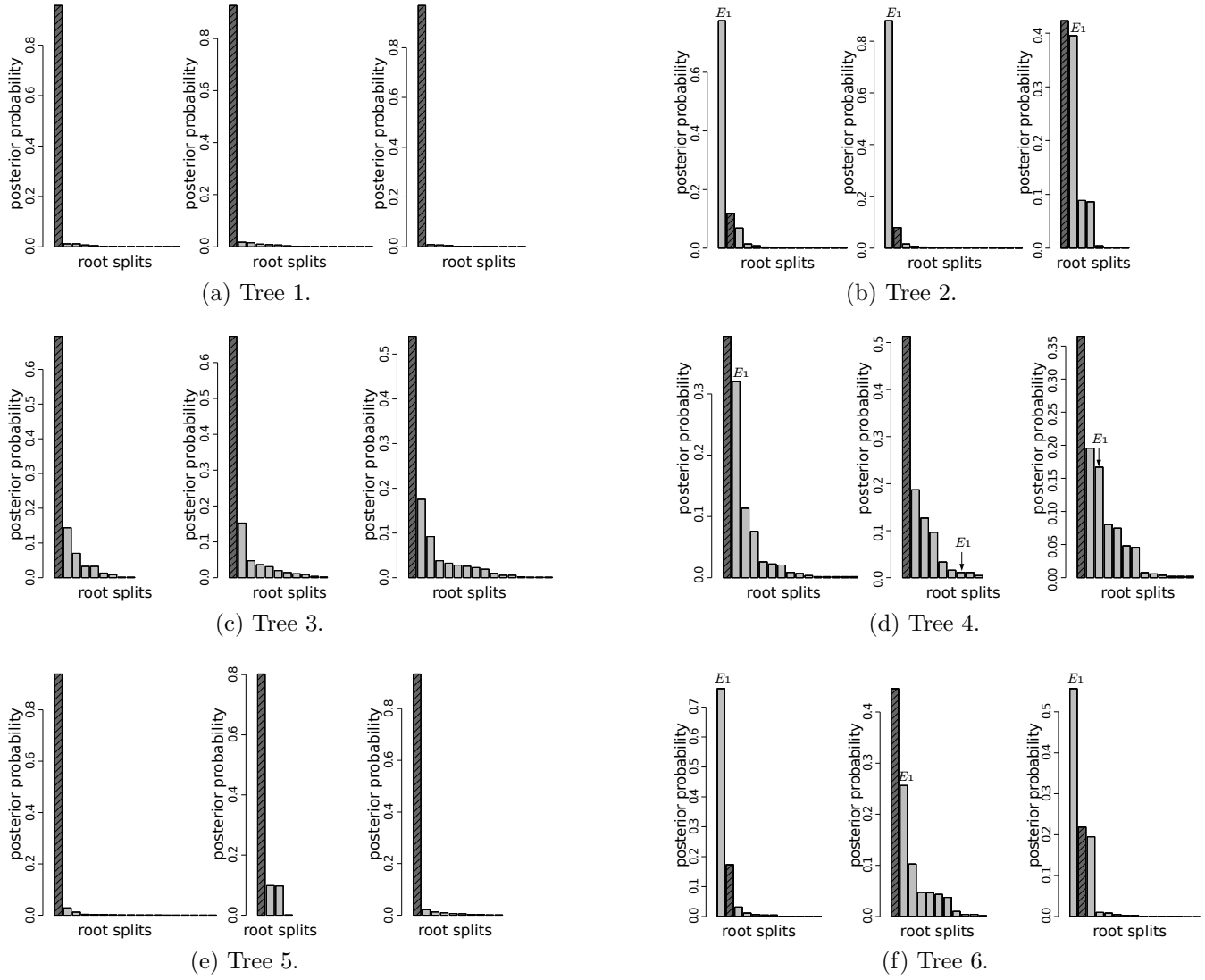
We use the NR model to simulate a rate matrix Q with $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)$, $\kappa = 2$ and $\sigma = 0.3$. In turn, this rate matrix is used to simulate three different alignments for each tree. These alignments are then analysed under the NR model with the Yule prior.

Tree 1: Tree 1 is rooted on the long branch E_1. Clearly the likelihood for data generated from this tree will support the correct placement of the root. Moreover, for the reasons expressed above, the prior will also support rooting on edge E_1. It is not surprising, therefore, that we find the posterior is very concentrated around the true root position (Figure 2a).

Tree 2: In Tree 2, the root is placed on the much shorter branch E_2, creating a fairly unbalanced unrooted topology with a long interior branch E_1. As such, data generated under this tree will favour the correct root position on edge E_2, but the prior will favour a root on branch E_1. This creates prior-likelihood conflict. As expected, we find that the posterior probability of the true root drops substantially in comparison to the analysis for Tree 1 and in two of the three analyses, the posterior offers more support to a root on edge E_1 (Figure 2b).

Tree 3: Tree 3 has the same rooted topology as Tree 1 but the root branch E_1 is now much shorter and the unrooted topology does not contain any long edges. As for Tree 1, prior-likelihood conflict does not arise but there is no longer such pronounced prior support for placement of the root on edge E_1. Nevertheless, we find that the posterior is still concentrated around the true root position (Figure 2c).

Tree 4: Tree 4 has the same rooted topology as Tree 2 but the long interior branch E_1 is



(a) Tree 1.

(b) Tree 2.

(c) Tree 3.

(d) Tree 4.

(e) Tree 5.

(f) Tree 6.

Figure 2: Posterior distribution of the root splits for three different alignments simulated for each of the six rooted trees according to Table 5 . Different bars on each plot represent different root splits ordered by posterior probabilities, with the highlighted bar representing the true root split. In the plots for Trees 2, 4 and 6, the split corresponding to a root on edge E_1 is also marked.

now shortened to 0.1. Although the Yule prior generally favours more balanced trees than Tree 4, the prior for branch lengths no longer offers overwhelming support to placement of the root on edge E_1. We find that the true root can now be recovered as the posterior mode (Figure 2d) but with less support than in the analysis for Tree 3.

Tree 5: Tree 5 has the same rooted topology as Trees 1 and 3, but the root edge E_1 has length 0.3, which lies between the corresponding values for Trees 1 and 3. As expected, we find that the true root is inferred as the posterior mode (Figure 2e), and the posterior is less (more) concentrated around the mode in comparison to the analysis of Tree 1 (Tree 3).

Tree 6: Tree 6 has the same rooted topology as Trees 2 and 4, but the internal edge E_1 has length 0.3, which lies between the corresponding values for Trees 2 and 4. The unrooted topology has a moderately long interior edge and the rooted topology is unbalanced, leading to some prior-likelihood conflict. We find that a root on edge E_1 sometimes receives more posterior support than the true root (Figure 2f), although, as expected, this effect is less pronounced than in the analysis for Tree 2.

This simulation experiment illustrates the sensitivity of root inferences to conflict between the prior and the likelihood. The effect of a mismatch in information about branch lengths is particularly noticeable. Given a particular unrooted topology, whilst the likelihood might support the presence of a long branch in the corresponding rooted tree, an Exp(10) prior does not, and therefore favours placement of the root on the long edge. Ideally constructing a more flexible prior that more explicitly models topology and branch lengths jointly will contribute to better root inference. However, given the absence of very long branches, our results show that the model is still able to extract information from the data about the root even in the face of prior-likelihood conflict.

Analysis of experimental data

Rooting the radiation of palaeopolyploid yeasts.—

We next investigated the performance of the NR and NR2 models on a real biological data set for which there is broad biological consensus on the root position (Byrne and Wolfe 2005; Hedtke et al. 2006). The lineage leading to *Saccharomyces cerevisiae* (brewer's yeast) and its relatives underwent a conserved whole-genome duplication (WGD) about 100 million years ago (Wolfe and Shields 1997; Kellis et al. 2004). Evidence for this WGD, in the form of duplicated genes and genomic regions, is shared by all post-WGD yeasts and defines the group as a clade from which the root of the *Saccharomycetales* is excluded (Figure 3) (Byrne and Wolfe 2005; http://ygob.ucd.ie 2015).

The root inferred through outgroup analysis separates a clade comprising *Eremothecium gossypii, Eremothecium cymbalariae, Kluyveromyces lactis, Lachancea kluyveri, Lachancea thermotolerans* and *Lachancea waltii* from the other species (Hedtke

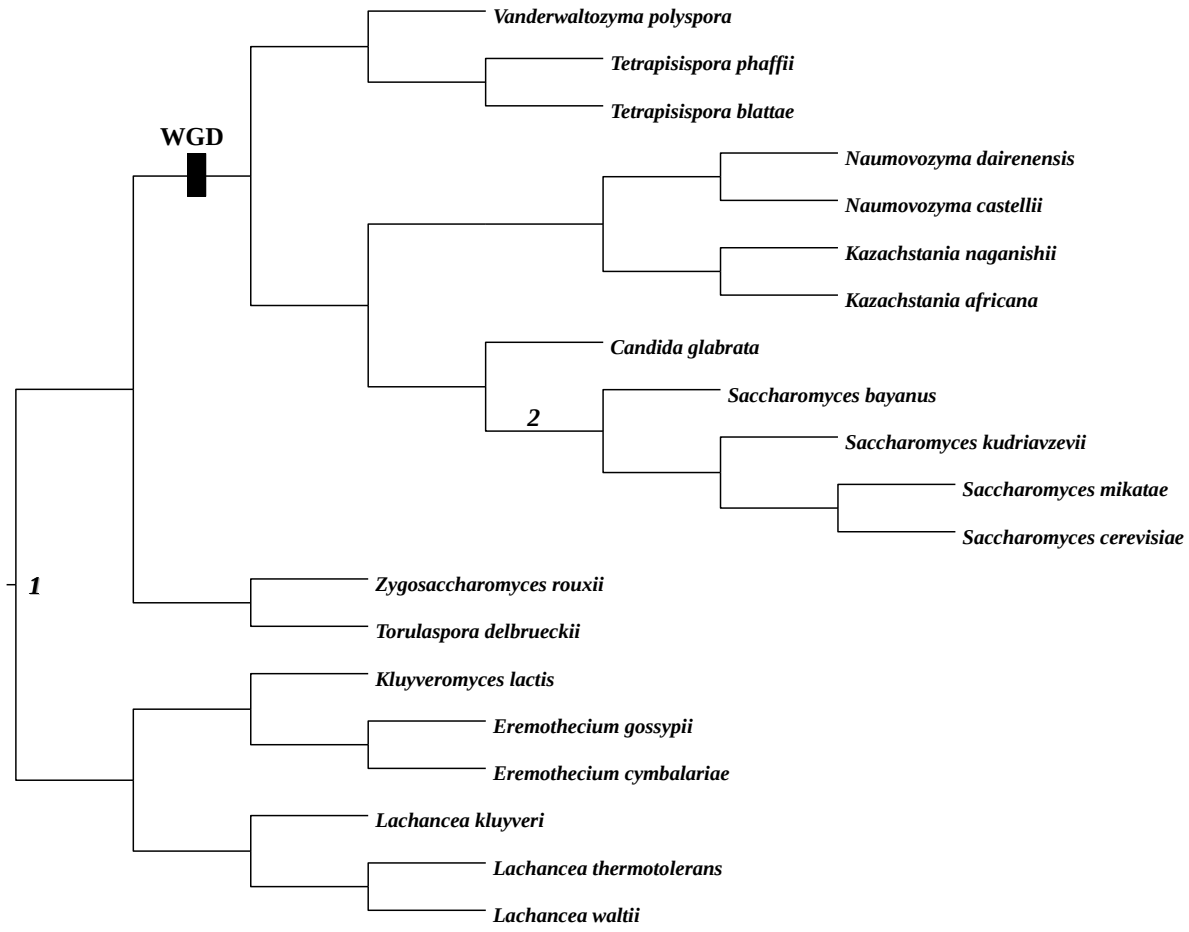
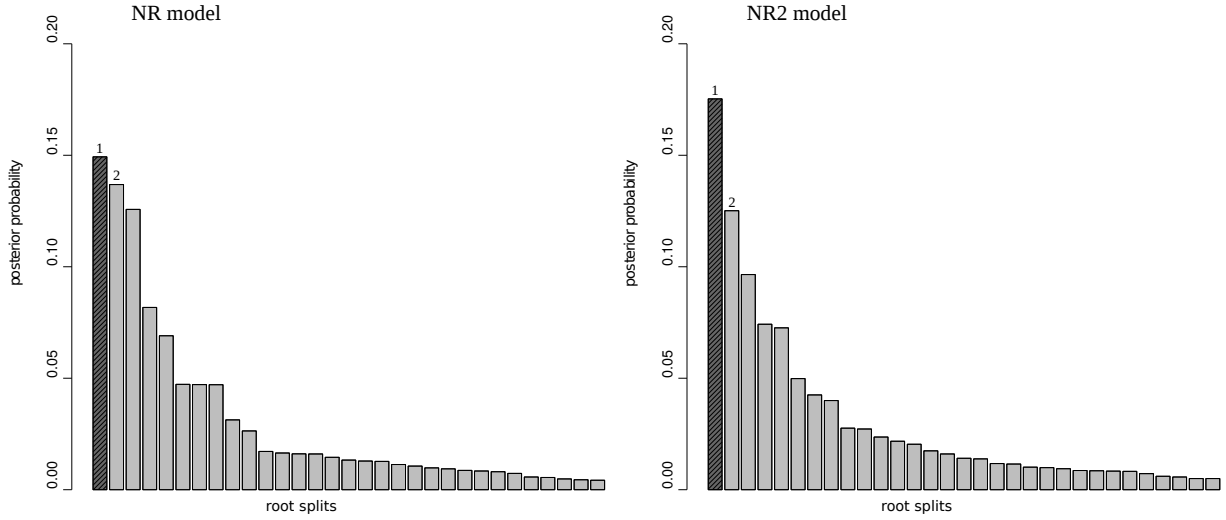


Figure 3: Rooted phylogeny of the palaeopolyploid yeasts supported by the whole-gene duplication analysis (not drawn to scale), reproduced from the YGOB website (Byrne and Wolfe 2005; http://ygob.ucd.ie 2015). The tree is rooted according to the outgroup method based on an analysis with the GTR+I+G model in a maximum likelihood framework (Hedtke et al. 2006). Roots 1 and 2 represent the two most plausible posterior root splits in the current analysis.

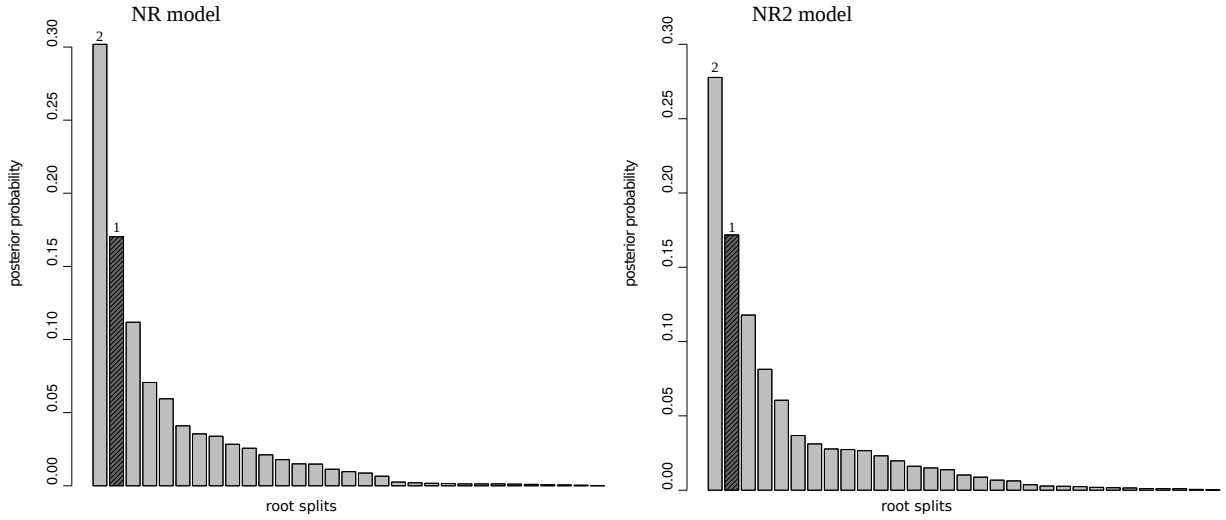
et al. 2006). We analysed an alignment of concatenated large and small subunit ribosomal DNA sequences for 20 yeast species, with a combined length of 4460 bp. The sequences were aligned with MUSCLE (Edgar 2004), and poorly aligned regions were detected and removed using TrimAl (Capella-Gutiérrez et al. 2009). We analysed this data set with the NR and NR2 models, using both the Yule prior and the structured uniform prior. In the analysis with the structured uniform prior, the root split supported by outgroup rooting (Hedtke et al. 2006) has the highest posterior probability (root 1 in Fig. 3) for both models. However, there is a substantial amount of uncertainty represented by the non-negligible posterior probabilities of the other root splits (Fig. 4a) and, for example, the second most plausible root is located within the post-WGD clade (root 2 in Fig. 3). This posterior uncertainty is also reflected in the sensitivity of the analysis to the topological prior: while the structured uniform prior recovered the root supported by the outgroup analysis with the highest posterior support, the Yule prior instead recovered this root with the second-highest support (Fig. 4b). The most plausible root inferred with the Yule prior is placed within the post-WGD clade (root 2 in Fig. 3) contradicting the WGD analysis.

The posterior for Huelsenbeck's I statistic is suggestive of a non-negligible degree of non-reversibility in the data (the posterior mean is 0.2 for the analysis with the NR model, 0.14 for the analysis with the NR2 model). In our simulations, larger values of I were generally required to infer the true root with high posterior probability. However, the support offered to the widely accepted outgroup root in this analysis shows that it is possible to extract useful root information in spite of the data suggesting only a modest degree of non-reversibility.

The unrooted topologies of the rooted majority rule consensus trees from the analyses with the two topological priors (Fig. 5) differ from that supported by the WGD analysis by the placement of *Vanderwaltozyma polyspora*. While the WGD analysis places it within the post-WGD clade, in our analysis this taxon is located within the pre-WGD clade. This result is consistent with our posterior inferences from fitting the HKY85 and GTR models. Interestingly, it is also consistent with the analysis performed with the site-heterogeneous CAT-GTR model (Lartillot and Philippe 2004) where *Vanderwaltozyma polyspora* is, again, excluded from the post-WGD clade (not shown). The placement of *Vanderwaltozyma polyspora* outside the WGD clade is surprising given that the genome of *Vanderwaltozyma polyspora* preserves evidence of having undergone WGD (Scannell et al. 2007). While this result requires further investigation, the similarity between the consensus trees obtained with the CAT-GTR model and with our non-reversible models suggests that the non-reversible models can not only extract meaningful information about the root position, but also capture information for inferring the unrooted topology. However, the minor mismatch of the topologies inferred in our analysis with that supported by WGD and outgroup analyses (Hedtke et al. 2006) confirms the presence of some features of the data that our models do not account for. For example, ribosomal RNA function depends on the molecule folding into a complex three-dimensional shape. Interactions among sites that are distant in the primary sequence, but close in the three dimensional structure, are likely to induce site-specific selective constraints that are not accounted for in our models.

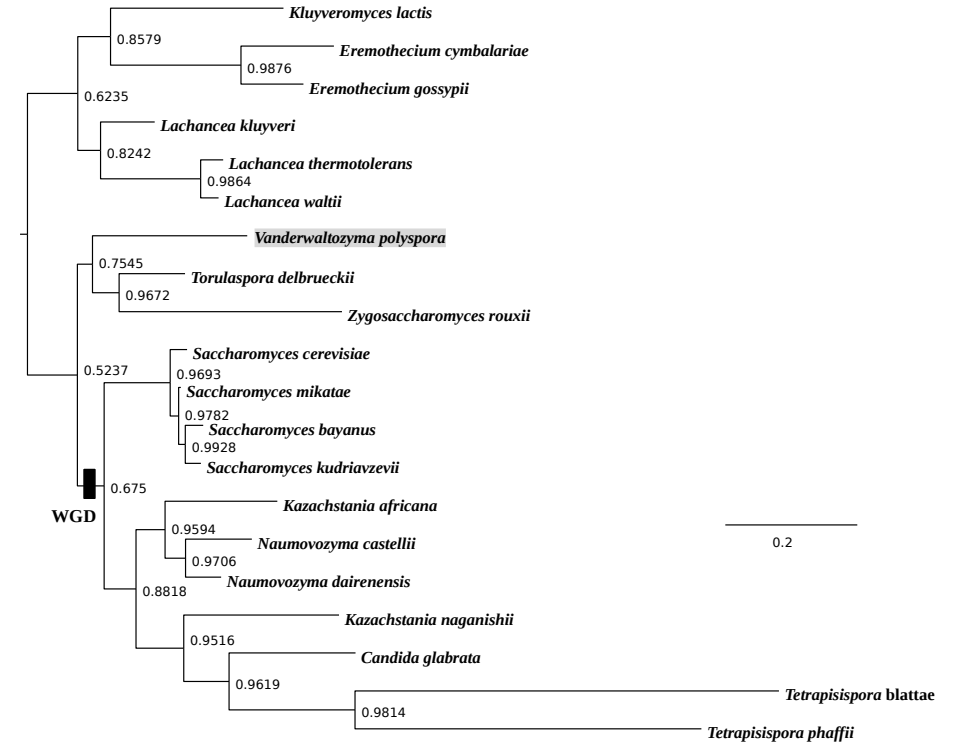


(a)

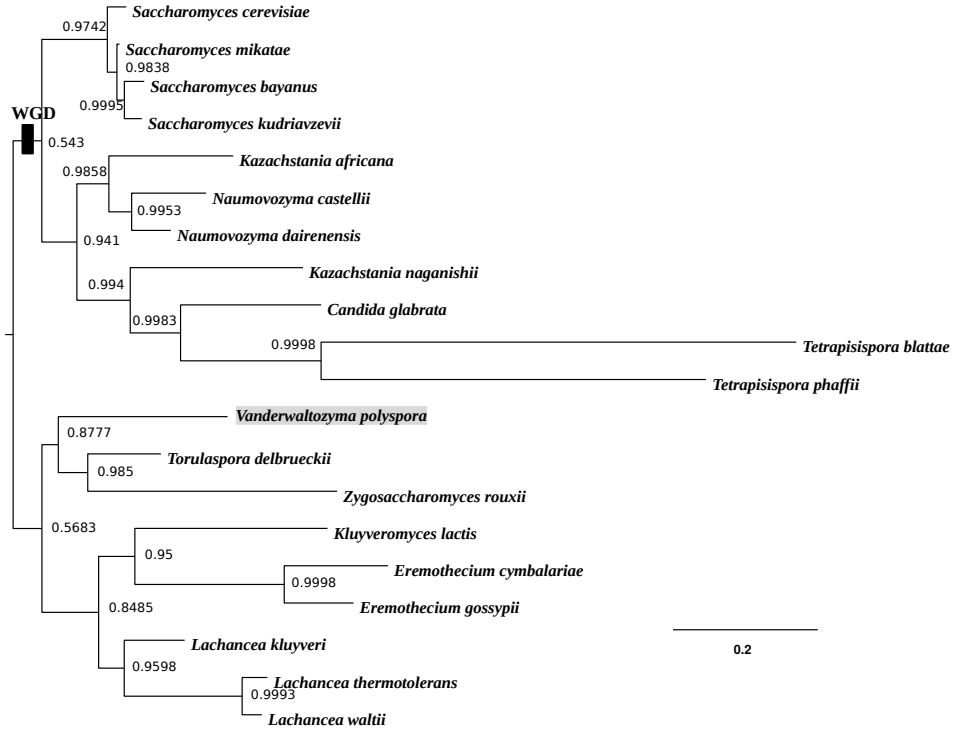


(b)

Figure 4: The posterior distribution of the root splits of the palaeopolyploid yeasts data set for both NR and NR2 models analysed (a) with the structured uniform prior and (b) with the Yule prior. Different bars on the plot represent different root splits on the posterior distribution of trees ordered by posterior probabilities (roots 1 and 2 are mapped in Figure 3). In (a), the analysis performed with the structured uniform prior, the root split supported by outgroup rooting (Hedtke at al. 2006) has the highest posterior probability (root 1, highlighted), while root 2 is placed within the post-WGD clade. In (b), the analysis performed with the Yule prior, the root split supported by outgroup rooting (Hedtke at al. 2006) has the second highest posterior probability (root 1, highlighted). The posterior modal root 2 is placed within the post-WGD clade.



(a)



(b)

Figure 5: Rooted majority rule consensus tree of the palaeopolyploid yeasts data set, inferred under the NR model using (a) the structured uniform prior and (b) the Yule prior, with the WGD event mapped. The analysis is based on the alignment of concatenated large and small subunit ribosomal DNA sequences for 20 yeast species, 4460 bp. The trees differ from that supported by the WGD analysis by the placement of *Vanderwaltozyma polyspora* (highlighted) within the pre-WGD clade.

In addition, it has been previously shown that failure to account for compositional heterogeneity can lead to inferring incorrect topologies with strong support (Foster 2004; Cox et al. 2008; Foster et al. 2009; Williams et al. 2012). Thus further refinement of the models, for instance, relaxing the stationarity assumption, might be necessary to improve the ability of the models to provide better insight into the evolution of palaeopolyploid yeasts.

It is worth noting that the root split on the majority rule consensus tree (Fig. 5b) does not match the marginal posterior modal root split (Fig. 4b). This happens because the consensus tree is a conditional summary, computed recursively from the leaves to the root, which depends upon the plausibility of subclades. On the other hand, the posterior over root splits is a marginal summary that averages over the relationships expressed elsewhere in the tree; see Appendix 2 for an illustrative example.

Analysis of the ribosomal tree of life.—

We have also applied the models to a data set for which there is still debate about the unrooted topology and root position: the ribosomal tree of life. Recall that the debates are centred on two hypotheses. According to the three-domains hypothesis, Archaea is monophyletic, sharing a common ancestor with Eukaryota (Woese 1990). The other hypothesis, called the eocyte hypothesis, suggests that Archaea is paraphyletic and Eukaryota originated from within Archaea (Lake 1988; Rivera and Lake 1992; Cox et al. 2008). Recent analyses of ribosomal RNA data have demonstrated that topological inferences can be sensitive to the choice of substitution model. When homogeneous models are used for the analysis they often recover the three-domains tree, while heterogeneous models generally recover the eocyte tree (Cox et al. 2008; Williams et al. 2012). In addition, there is also external evidence for the eocyte hypothesis. For example, newly discovered archaeal species whose genomes encode many eukaryote-specific features, provide additional support for the eocyte hypothesis (Spang et al. 2015).

Here we analysed aligned concatenated large and small subunit ribosomal RNA sequences from archaeal, bacterial and eukaryotic species (36 taxa, 1734 sequence positions), including the recently discovered archaeal groups: Thaumarchaeota, Aigarchaeota and Korarchaeota. These new groups are closely related to Crenarchaeota and together they form the so-called TACK superphylum (Guy and Ettema 2011; Kelly et al. 2011; Williams et al. 2012; Lasek-Nesselquist and Gogarten 2013). Previous analysis of this data set performed with the CAT-GTR model recovered an eocyte topology (Williams et al. 2012). Fitting the simpler HKY85 and GTR models also support this hypothesis. However these analyses were not able to infer the root because they used only reversible rate matrices in stationary substitution models. We analysed these data with both the NR and NR2 models using both the Yule prior and the structured uniform prior. In all cases we recovered the eocyte topology with similar posterior support (Fig. 6). The analysis with the Yule prior assigned high posterior support to two roots splits (Fig. 7a) - one on the branch leading to Bacteria (root 1 in Fig. 6), the other within Bacteria, on the branch leading to *Rhodopirellula baltica* (root 2 in Fig. 6). This inference is in accord with current biological opinion about the root of the tree of life, which places the root either on

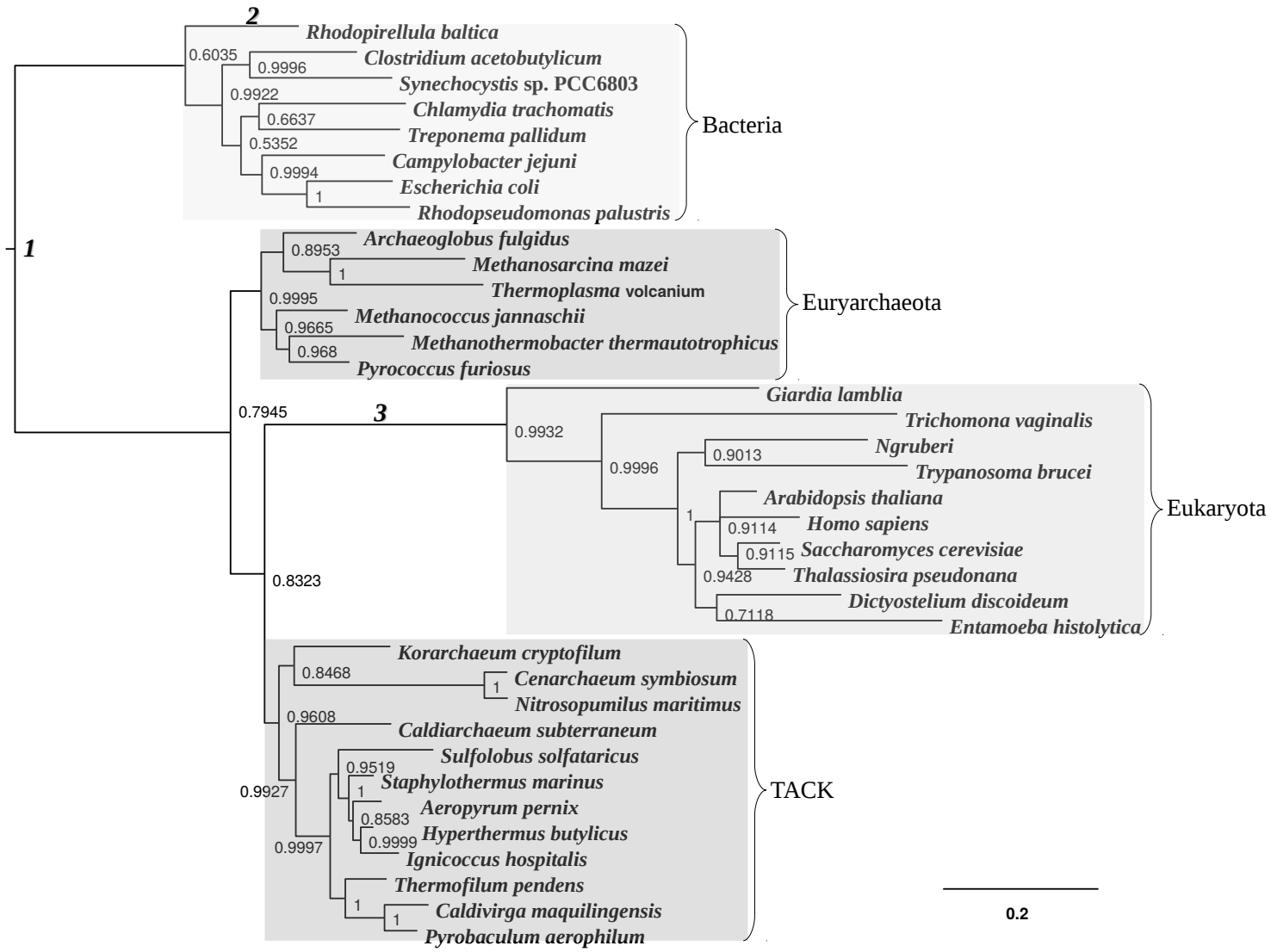


Figure 6: Rooted majority rule consensus tree for the tree of life data set, inferred under the NR model using the Yule prior. The tree supports the eocyte hypothesis by placing Eukaryota within Archaea, as a sister group to the TACK superphylum. Roots 1, 2 and 3 are the root splits having the highest posterior support in the current analysis. Posterior support for these root splits is shown in Figure 7.

the branch leading to Bacteria, or within Bacteria (Baldauf et al. 1996; Cavalier-Smith 2006; Skophammer et al. 2007; Hashimoto and Hasegawa 1996). However, in the analysis performed with the structured uniform prior, the support for the root within Bacteria decreased and that for the the root on the bacterial branch increased (Fig. 7b). This analysis illustrates the sensitivity of the inference to the choice of topological prior, and confirms the importance of the choice of prior in Bayesian phylogenetics. The posterior

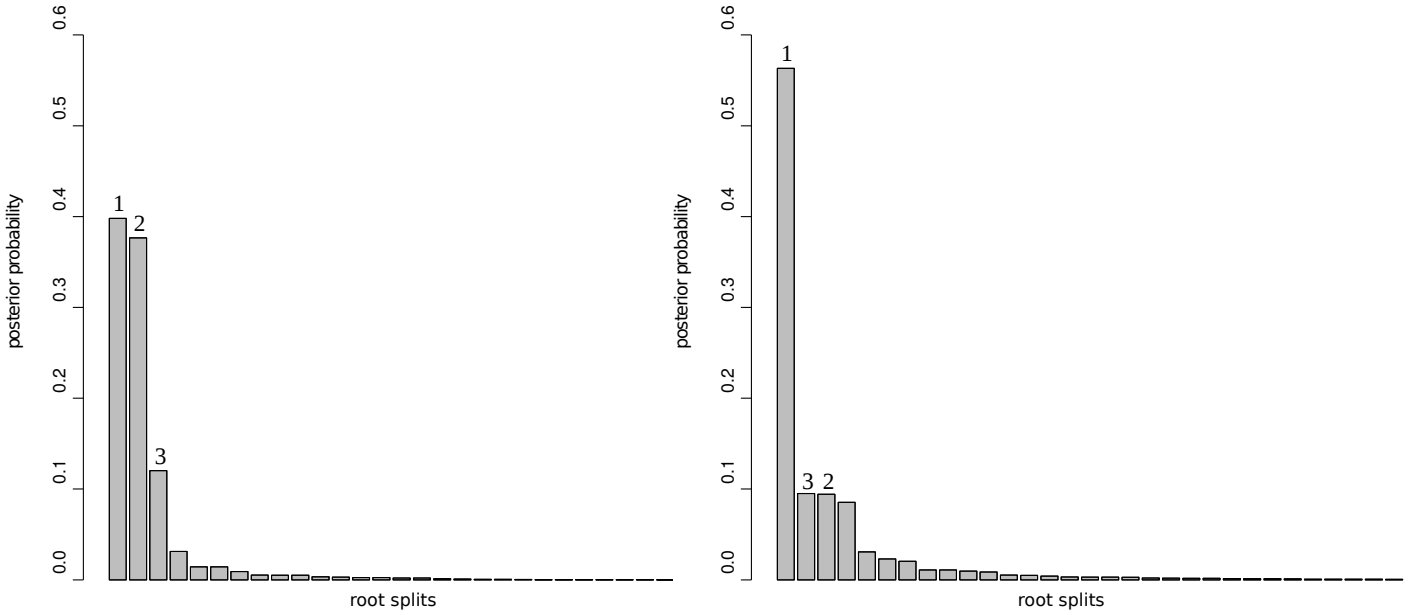


Figure 7: The posterior distribution of the root splits of the tree of life data set for the NR model analysed with (a) the Yule prior; and (b) with the structured uniform prior. Different bars on the plot represent different root splits on the posterior distribution of trees (ordered by posterior probabilities). The root split on the branch leading to Bacteria has the highest posterior probability (root 1). Root 2 is placed within Bacteria (on the branch leading to *Rhodopirellula baltica*) and root 3 is placed on the branch leading to Eukaryota (the roots are mapped in Figure 6).

mean of the Huelsenbeck's I statistic is 0.18 for the analysis with the NR model and 0.17 for the analysis with the NR2 model. Again, this is suggestive of a moderate degree of non-reversibility in the data. Therefore, modelling other features of the data that also provide root information could make a valuable contribution to the inference.

Discussion

We presented two hierarchical non-reversible models for inferring rooted phylogenetic trees. The non-reversibility of both models is achieved by applying a stochastic perturbation to the rate matrix of a reversible model. This perturbation makes the likelihood dependent on the position of the root, enabling us to infer the root directly from the sequence alignment. In the first model (the NR model) we use only one variation component and perform a log-normal perturbation on the space of all possible rate matrices. In contrast, the second model (the NR2 model) utilises two variation components and the perturbation is performed on the space of reversible and non-reversible rate matrices separately. This separation allows us to judge the extent of the different types of perturbation.

The results on the simulated data with different levels of non-reversibility show that the correct root can be recovered with greater posterior support when the degree of non-reversibility in the data generating model is larger. We also investigated the robustness of posterior root inferences to situations where information from the prior and data are in conflict. Given a particular unrooted topology, our Yule prior for rooted trees and Exp(10) prior for branch lengths offers most support to balanced trees with short edges. Our simulations show that we can still recover the true root in the posterior when the data generating tree is unbalanced or the associated unrooted topology contains a long edge. However, when this edge is very long, it can mislead the root inference.

We applied our models to two biological data sets. These analyses agree with our simulations in suggesting that our non-reversible models can recover useful rooting information, this time from real biological sequence alignments. The analyses of both the yeast and tree of life data sets recover the widely agreed root. However, both data sets show some prior sensitivity, even though the two topological priors (the Yule prior and the structured uniform prior) share similar features. In order to investigate this issue we computed a log Bayes factor (Kass and Raftery 1995) to compare the Yule prior (Y) with the structured uniform prior (S) for both examples with real data. Although usually used to compare models, the Bayes factor really compares prior-likelihood combinations and so can also be used to assess which of the two priors is most consistent with the data. The log Bayes factor for the yeasts data set is $\log B_{YS} = 2.27$ suggesting that there is evidence against the structured uniform prior, however, the evidence is not strong. The log Bayes factor for the tree of life data set is $\log B_{YS} = 0.12$ suggesting that there is no difference between the priors. Therefore the more noticeble prior sensitivity in the analysis of the yeasts data set is likely to be due to the greater difference in consistency between the data and each of the two priors.

Although Huelsenbeck's I statistic provides evidence of a non-negligible degree of non-reversibility in both biological data sets, the analyses display high levels of posterior uncertainty. This suggests that the information about the root may be obscured by other signals that are not accounted for by our current models. For instance, our models assume the evolutionary process is stationary. If this was true then the empirical composition of the four nucleotides would be roughly the same for all taxa in the alignment. However, this is often not the case in experimental data (Foster 2004; Cox et al. 2008). Notably, this assumption is violated for the tree of life data set where the empirical GC content ranged from 41% for *Entamoeba histolytica* to 69% for *Giardia lamblia*. The models may therefore benefit from further development, for example to model the non-stationarity of the process. Nonetheless, our findings illustrate that our non-reversible models NR and NR2 can be useful to infer the root position from real biological data sets.

Materials and methods

We work within the Bayesian paradigm and base our inferences on the posterior distribution of the unknowns in the model. According to Bayes theorem, the posterior

distribution is proportional to the prior times the likelihood. For the NR model, for example, the posterior distribution factorises as

$$\begin{aligned}\pi(\boldsymbol{\pi}, \kappa, \sigma, Q, \alpha, \boldsymbol{\ell}, \tau|\text{Data}) &\propto \pi(Q|\boldsymbol{\pi}, \kappa, \sigma) \\ &\times \pi(\boldsymbol{\pi}, \kappa, \sigma, \alpha, \boldsymbol{\ell}, \tau) \times \pi(\text{Data}|Q, \alpha, \boldsymbol{\ell}, \tau).\end{aligned} \tag{3}$$

This distribution is analytically intractable and so we build up a numerical approximation by sampling from it using Markov chain Monte Carlo (MCMC) methods, specifically a Metropolis-within-Gibbs sampling scheme. In the remainder of this section, we first describe the calculation of the likelihood function, before outlining details of the MCMC algorithm. Finally, we provide practical details of the application of this algorithm to the analyses presented earlier in the *Results* section.

Likelihood

The likelihood function summarises the information available from the data about the unknowns in the model including the phylogenetic tree. Since we assume that alignment sites evolve independently of each other, the likelihood can be expressed as a product of the likelihoods of the n individual sites of the alignment. If we denote $\boldsymbol{\theta}$ to be the parameters of the substitution process, the likelihood takes the form

$$\pi(\text{Data}|\boldsymbol{\theta}, \alpha, \boldsymbol{\ell}, \tau) = \prod_{i=1}^{n} \pi(D_i|\boldsymbol{\theta}, \alpha, \boldsymbol{\ell}, \tau),$$

where D_i is the column of nucleotides at site i. The probability of the data at a site i is given by

$$\pi(D_i|\boldsymbol{\theta}, \alpha, \boldsymbol{\ell}, \tau) = \sum_{X} \pi_{X(root)} \prod_{\text{edges } \ell=(v,w)} p_{X(v),X(w)}(\ell)$$

where v and w are the vertices at the two ends of edge ℓ and $X(u)$ denotes the nucleotide at a vertex u. The sum is taken over all functions X from the vertices to Ω such that $X(u)$ matches data $D_i(u)$ for all leaf vertices u. We assume a stationary model and so take the probability at the root $\pi_{X(root)}$ to be $\pi_{Q,X(root)}$, which comes from $\boldsymbol{\pi}_Q$, the theoretical stationary distribution associated with Q (note that this is not the same as $\boldsymbol{\pi}$, the stationary distribution of the underlying HKY85 model).

MCMC algorithm

NR model.— For the NR model, the posterior distribution for the unknowns in the model was summarised through equation (3). At each iteration of the MCMC algorithm the following steps are performed:

(a) update the parameters of the substitution model $(\boldsymbol{\pi}, \kappa, \sigma, Q, \alpha)$;

(b) update the branch lengths $\boldsymbol{\ell}$ and the rooted topology τ.

In step (a) we update the parameters using a Dirichlet random walk proposal for $\boldsymbol{\pi}$ and log-normal random walk proposals for the other parameters. Move (b) consists of a series of Metropolis-Hastings steps to update each branch length one at a time using a log-normal random walk proposal and then updating the rooted topology and branch lengths (in a joint move) through three types of proposal: nearest-neighbour interchange (NNI), sub-tree prune and regraft (SPR), and a proposal that moves the root; see Heaps et al. (2014) for complete details of all three moves.

NR2 model.— Here the posterior distribution of the unknowns takes the form

$$\begin{aligned}\pi(\boldsymbol{\pi}, \kappa, \sigma_R, \sigma_N, \boldsymbol{\epsilon}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau|\text{Data}) \\ \propto \pi(\boldsymbol{\pi}, \kappa, \sigma_R, \sigma_N, \boldsymbol{\epsilon}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau) \\ \times \pi(\text{Data}|\boldsymbol{\pi}, \kappa, \boldsymbol{\epsilon}, \boldsymbol{\eta}, \alpha, \boldsymbol{\ell}, \tau)\end{aligned}$$

and an analogous Metropolis-within-Gibbs algorithm is used to generate posterior samples.

MCMC implementation

In the *Results* section, all results were based on (almost) un-autocorrelated posterior samples of size 5K. These samples were obtained by running the MCMC algorithm for at least 1000K iterations, discarding at least 500K iterations as burn-in and then thinning by taking every 100th iterate to remove autocorrelation. Convergence was diagnosed using the procedure described in Heaps et al. (2014). This involved initialising two MCMC chains at different starting points and graphically comparing the chains through properties based on model parameters and the relative frequencies of sampled clades. In all cases, the graphical diagnostics gave no evidence of any lack of convergence. The MCMC inferential procedures are programmed in Java and a software implementation can be found in the Supplementary Materials.

Appendix 1

The two-stage perturbation relies upon the underlying geometry of the space of Markov rate matrices, and is achieved in the following way. We work on a log-scale element-wise with all matrices, ignoring diagonal elements. The set of all possible 4×4 rate matrices M is therefore identified with $\mathbb{R}^{12}$ which we equip with the standard inner product. The set of HKY85 matrices and GTR matrices form nested sub-sets of M. Recall that working element-wise on a log-scale, the off-diagonal elements of the rate matrix of the NR model can be expressed as, for $i \neq j$

$$\log q_{ij} = \log q_{ij}^H + \epsilon_{ij}, \tag{4}$$

where the ϵ_{ij} are independent $N(0, \sigma^2)$ quantities. The element-wise log of the HKY85 matrix Q^H in equation (4) is

$$\begin{aligned}\log q_{ij}^H &= \tilde{\kappa}(\mathbf{e}_1\mathbf{e}_2^T + \mathbf{e}_2\mathbf{e}_1^T + \mathbf{e}_3\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_3^T) + \sum_{i=1}^{4} \tilde{\pi}_i \mathbf{s}\,\mathbf{e}_i^T \\ &= \tilde{\kappa}(\mathbf{e}_1\mathbf{e}_2^T + \mathbf{e}_2\mathbf{e}_1^T + \mathbf{e}_3\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_3^T) + \sum_{i=1}^{3} \tilde{\pi}_i \mathbf{s}\,\mathbf{e}_i^T \\ &+ \log\left(1 - e^{\tilde{\pi}_1} - e^{\tilde{\pi}_2} - e^{\tilde{\pi}_3}\right) \mathbf{s}\,\mathbf{e}_4^T\end{aligned} \tag{5}$$

where $(\tilde{\pi}_1, \tilde{\pi}_2, \tilde{\pi}_3, \tilde{\pi}_4)$=$(\log \pi_A, \log \pi_G, \log \pi_C, \log \pi_T)$, $\tilde{\kappa} = \log \kappa$, $\mathbf{e}_i$ is the i-th standard basis vector of $\mathbb{R}^4$ and $\mathbf{s} = (1, 1, 1, 1)^T$. By differentiating (5) with respect to the parameters $\tilde{\pi}_1, \tilde{\pi}_2, \tilde{\pi}_3$ and $\tilde{\kappa}$ we obtain 4 linearly independent vectors in M that are locally tangent to the sub-set of HKY85 matrices at Q^H, and we denote these V_1, V_2, V_3, V_4. (Differentiating with respect to $\tilde{\pi}_4$ gives a tangent vector contained in the span of V_1, V_2, V_3.) The tangent vectors in M correspond to the 4×4 matrices

$$V_i = \mathbf{s}\,\mathbf{e}_i^T - \exp(\tilde{\pi}_i - \tilde{\pi}_4)\mathbf{s}\,\mathbf{e}_4^T \quad \text{for } i = 1, 2, 3,$$

and

$$V_4 = \mathbf{e}_1\mathbf{e}_2^T + \mathbf{e}_2\mathbf{e}_1^T + \mathbf{e}_3\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_3^T.$$

The element-wise log of the general GTR matrix is

$$\sum_{i=1}^{4} \tilde{\pi}_i \mathbf{s}\,\mathbf{e}_i^T + \sum_{i<j} \tilde{\rho_{ij}} \left(\mathbf{e}_i\mathbf{e}_j^T + \mathbf{e}_j\mathbf{e}_i^T\right),$$

where $\tilde{\rho}_{ij}$ is the log of the exchangeability parameter ρ_{ij}. By considering the derivatives with respect to the $\tilde{\rho}_{ij}$ parameters, it can be seen that the the following vectors lie in the tangent space to the GTR matrices at Q^H:

$$\begin{aligned}V_5 &= \left(\mathbf{e}_1\mathbf{e}_2^T + \mathbf{e}_2\mathbf{e}_1^T\right) - \left(\mathbf{e}_3\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_3^T\right), \\ V_6 &= \left(\mathbf{e}_1\mathbf{e}_3^T + \mathbf{e}_3\mathbf{e}_1^T\right) + \left(\mathbf{e}_2\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_2^T\right), \\ V_7 &= \left(\mathbf{e}_1\mathbf{e}_3^T + \mathbf{e}_3\mathbf{e}_1^T\right) - \left(\mathbf{e}_2\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_2^T\right), \\ V_8 &= \left(\mathbf{e}_1\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_1^T\right) + \left(\mathbf{e}_2\mathbf{e}_3^T + \mathbf{e}_3\mathbf{e}_2^T\right), \\ V_9 &= \left(\mathbf{e}_1\mathbf{e}_4^T + \mathbf{e}_4\mathbf{e}_1^T\right) - \left(\mathbf{e}_2\mathbf{e}_3^T + \mathbf{e}_3\mathbf{e}_2^T\right).\end{aligned}$$

The vectors $V_1, V_2, \ldots, V_9$ are linearly independent by construction. Standard linear algebra can be used to extend this to a basis $V_1, \ldots, V_{12}$ of $\mathbb{R}^{12}$.

Next, the QR factorisation algorithm is applied to the 12×12 matrix with columns $V_1, \ldots, V_{12}$ to obtain an orthonormal basis of tangent vectors $W_1, \ldots, W_{12}$ that is used to

perturb Q^H. First, Q^H is perturbed using $\nu_1, \ldots, \nu_5$ to obtain a GTR matrix Q^R where, for $i \neq j$

$$\log q_{ij}^R = \log q_{ij}^H + \sum_{k=5}^{9} \nu_{k-4} W_{kij},$$

and the ν_k are independent $N(0, \sigma_R^2)$ and W_{kij} is the (i, j)-th element of the 4×4 matrix corresponding to W_k. The choice of basis $W_1, \ldots, W_{12}$ ensures that this perturbation is locally orthogonal to the sub-set of HKY85 matrices, and that the perturbation is otherwise isotropic within the sub-set of GTR matrices. The second stage perturbs Q^R into the space of non-reversible rate matrices using η_1, η_2, η_3: for $i \neq j$

$$\log q_{ij} = \log q_{ij}^R + \sum_{k=10}^{12} \eta_{k-9} W_{kij},$$

and the η_k are independent $N(0, \sigma_N^2)$ quantities. This perturbation is locally perpendicular to the sub-set of GTR matrices in M. The equation determines the off-diagonal elements of the non-reversible rate matrix Q, while the diagonal elements are fixed in order to make the row sums zero. The size of the perturbation variance σ_R^2 can be thought of as representing the extent to which the rate matrix Q departs from the class of HKY85 models remaining within the class of reversible models, while σ_N^2 represents the extent to which Q departs from being reversible.

Appendix 2

The root on the majority rule consensus tree and the mode of the posterior distribution for root splits are different point summaries of the posterior distribution for root positions. Both can be approximated from posterior samples of rooted topologies but they need not coincide. For example, suppose the posterior output comprises the following five trees:

Tree 1:	((A,B),(((E,F),D),C));
Tree 2:	(((A,B),C),((E,F),D));
Tree 3:	((((A,B),C),D),(E,F));
Tree 4:	(((((A,B),C),D),E),F);
Tree 5:	((A,B),(((E,F),D),C));

The clade (A, B) appears on all the trees, and so is included in the consensus tree with probability one. Similarly, the clade (A, B, C) appears on three trees (Tree 2, Tree 3 and Tree 4), and so appears in the consensus tree with support 0.6. Continuing in this fashion, the consensus tree is completed by incorporating the clades (E, F) and (D, E, F) that appear with support 0.8 and 0.6 respectively. Hence, the root position on the consensus tree (displayed in Figure 8) separates the taxa A, B, C from D, E, F. On the

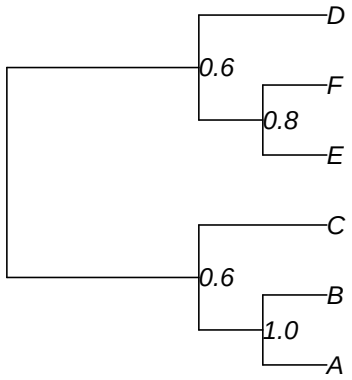


Figure 8: Majority rule consensus tree for illustrative example.

Table 6: Posterior for root splits in illustrative example.

Root split	**Count**	**Probability**
(A, B) : (C, D, E, F)	2	0.4
(A, B, C) : (D, E, F)	1	0.2
(E, F) : (A, B, C, D)	1	0.2
(F) : (A, B, C, D, E)	1	0.2

other hand, the posterior for root splits is given in Table 6. Clearly the posterior modal root split is (A, B) : (C, D, E, F) which does not match the root split (A, B, C) : (D, E, F) on the consensus tree.

Acknowledgments

This work was supported by the European Research Council (Advanced Investigator Award, grant number ERC-2010-AdG-268701, supporting S.C., S.E.H., T.A.W. and T.M.E.); and the Wellcome Trust (Program Grant, number 045404, to T.M.E.)

M. E. Alfaro and M. T. Holder. The posterior and the prior in Bayesian phylogenetics. *Annu. Rev. Ecol. Evol. Syst.*, 37:19–42, 2006.

S. L. Baldauf, J. D. Palmer, and W. Ford Doolittle. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl. Acad. Sci. USA*, 93:7749–7754, 1996.

J. Bergsten. A review of long-branch attraction. *Cladistic*, 21:163–193, 2005.

S. Blanquart and N. Lartillot. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.*, 23(11): 2058–2071, 2006.

J. R. Brown and W. F. Doolittle. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl. Acad. Sci. USA*, 92: 2441–2445, 1995.

K. P. Byrne and K. H. Wolfe. The yeast gene order browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research*, 15: 1456–1461, 2005.

S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25:1972–1973, 2009.

T. Cavalier-Smith. Rooting the tree of life by transition analyses. *Biology Direct*, 1:19–19, 2006.

C. J. Cox, P. J. Foster, R. P. Hirt, S. R. Harris, and T. M. Embley. The archaebacterial origin of eukaryotes. *PNAS*, 51:20356–20361, 2008.

J. Dutheil and B. Boussau. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol.*, 28:255, 2008.

R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32:1792–1797, 2004.

T. M. Embley and W. Martin. Eukaryotic evolution, changes and challenges. *Nature*, 440: 623–630, 2006.

J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Biol*, 27:401–410, 1978.

P. G. Foster. Modeling compositional heterogeneity. *Syst. Biol*, 53(3):485–495, 2004.

P. G. Foster, C. J. Cox, and T. M. Embley. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Phil. Trans. R. Soc. B*, 364: 2197–2207, 2009.

N. Galtier and M. Gouy. Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.*, 15:871–879, 1998.

L. Guy and T. J. G. Ettema. The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends in Microbiology*, 19:580–587, 2011.

M. Hasegawa, H. Kishino, and T. Yono. Dating of human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.

T. Hashimoto and M. Hasegawa. Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1α/Tu and 2/G. *Adv. Biophys*, 32:73–120, 1996.

S. E. Heaps, T. M. W. Nye, R. J. Boys, T. A. Williams, and T. M. Embley. Bayesian modelling of compositional heterogeneity in molecular phylogenetics. *Stat. Appl. Genet. Mol. Biol.*, 1:1–21, 2014.

S. M. Hedtke, T. M. Townsend, and D. M. Hillis. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst. Biol*, 55:522–529, 2006.

B. R. Holland, D. Penny, and M. D. Hendy. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock - a simulation study. *Syst. Biol*, 52:229–238, 2003.

http://ygob.ucd.ie. Yeast Gene Order Browser. `http://ygob.ucd.ie//`, 2015. [Online; accessed 1-Jan-2015].

J. P. Huelsenbeck, J. P. Bollback, and A. M. Levine. Inferring the root of a phylogenetic tree. *Syst. Biol.*, 51(1):32–43, 2002.

N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. USA*, 86:9355–9359, 1989.

V. Jayaswal, F. Ababneh, L. S. Jermiin, and J. Robinson. Reducing model complexity of the general Markov model of evolution. *Mol. Biol. Evol.*, 28(11):3045–3059, 2011.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

M. Kellis, B. W. Birren, and E. S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428:617–624, 2004.

S. Kelly, B. Wickstead, and K. Gull. Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. *Proc. R. Soc. Lond. B.*, 278:1009–1018, 2011.

J. A. Lake. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature*, 331:184–186, 1988.

N. Lartillot and H. Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, 21:1095–1109, 2004.

E. Lasek-Nesselquist and J. P. Gogarten. The effects of model choice and mitigating bias on the ribosomal tree of life. *Mol. Biol. Evol.*, 69:17–38, 2013.

D. Penny. Criteria for optimising phylogenetic trees and the problem of determining the root of a tree. *Mol. Biol. Evol.*, 8:95–116, 1976.

M. C. Rivera and J. A. Lake. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science*, 257:74–76, 1992.

D. R. Scannell, A. C. Frank, G. C. Conant, K. P. Byrne, M. Woolfit, and K. H. Wolfe. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *PNAS*, 104:8397–8402, 2007.

R. G. Skophammer, J. A. Servin, C. W. Herbold, and J. A. Lake. Evidence for a gram-positive, eubacterial root of the tree of life. *Mol. Biol. Evol.*, 24:1761–1768, 2007.

A. Spang, J. H. Saw, S. L.Jørgensen, K. Zaremba Niedzwiedzka, J. Martijn, A. E. Lind, R. van Eijk, C. Schleper, L. Guy, and T. J. G. Ettema. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, 521:173–179, 2015.

F. Squartini and P. F. Arndt. Quantifying the stationarity and time reversibility of the nucleotide substitution process. *Mol. Biol. Evol.*, 25:2525–2535, 2008.

M. Steel and A. McKenzie. Properties of phylogenetic trees generated by Yule-type speciation models. *Mathematical Biosciences*, 170:91–112, 2001.

S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *American Mathematical Society*, 17:57–86, 1986.

N. J. Tourasse and M. Gouy. Accounting for evolutionary rate variation among sequence sites consistently changes universal phylogenies deduced from rRNA and protein-coding genes. *Molecular Phylogenetics and Evolution*, 13:159–168, 1999.

T. A. Williams, P. G. Foster, T. M. W. Nye, C. J. Cox, and T. M. Embley. A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc. R. Soc.*, B 279: 4870–4879, 2012.

T. A. Williams, P. G. Foster, C. J. Cox, and T. M. Embley. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, 504:231–236, 2013.

C. R. Woese. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA*, 87:4576–4579, 1990.

K. H. Wolfe and D. C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–713, 1997.

M. D. Woodhams, J. Fernández-Sánchez, and J. G. Sumner. A new hierarchy of phylogenetic models consistent with heterogeneous substitution rates. *Syst. Biol.*, 64(4): 638–650, 2015.

Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10:1396–1401, 1993.

Z. Yang. Maximum-likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39:306–314, 1994.

Z. Yang and B. Rannala. Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst. Biol*, 53:455–470, 2005.

Z. Yang and D. Roberts. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol. Biol. Evol.*, 12:451–458, 1995.